ISSN 0104-0499

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

DEs-UFSCar e SME-ICMC-USP

DISTRIBUIÇÃO KUMARASWAMY-EXPONENCIAL PARA ANÁLISE DE DADOS DE SOBREVIVÊNCIA

Amanda Morales Eudes
Vera Lucia D. Tomazella
Cirdêmia Costa Feitosa

RELATÓRIO TÉCNICO

TEORIA E MÉTODO – SÉRIE A

Maio/2014
nº 260

Distribuição Kumaraswamy-Exponencial para Análise de Dados de Sobrevivência

Amanda Morales Eudes* Vera Lucia D. Tomazella †

Cirdêmia Costa Feitosa‡

6 de maio de 2014

Resumo

Na literatura, diversas distribuições conhecidas são utilizadas para acomodar dados de tempos de falha, porém, grande parte destas distribuições não é capaz de acomodar taxas de falha não monótonas. Kumaraswamy (1980) propôs uma nova distribuição de probabilidade capaz de acomodar tais taxas e, baseada nela, mais recentemente Cordeiro e de Castro (2009) propuseram uma nova família de distribuições generalizadas, a Kumaraswamy-Generalizada (Kum-G). Esta distribuição além de ser flexível, contém distribuições com funções de risco unimodal e em forma de banheira, como mostrado por Pascoal *et al*. (2011). Neste artigo, nós apresentamos a distribuição Kumaraswamy-Exponencial (Kum-Exp) para analisar dados de tempo de vida dos indivíduos em risco, sendo que este modelo é caso particular da família de distribuições Kum-G. Algumas propriedades desta distribuição serão apresentadas, assim como o método adequado de estimação para os parâmetros do modelo, de forma clássica e também bayesiana. A nova distribuição é ilustrada com dois conjuntos de dados da literatura.

Palavras-Chave: Análise de sobrevivência;Distribuição Kumaraswamy-G; Distribuição Kumaraswamy-Exponencial; abordagem bayesiana.

*Departamento de Estatística, Centro de Ciências Exatas e Tecnológicas, Universidade Federal de São Carlos, Caixa Postal 676, CEP: 13.565-905, São Carlos, São Paulo, Brasil, E-mail: *amanda_eudes@hotmail.com*

†Departamento de Estatística, Centro de Ciências Exatas e Tecnológicas, Universidade Federal de São Carlos, Caixa Postal 676, CEP: 13.565-905, São Carlos, São Paulo, Brasil, E-mail: *vera@ufscar.br*

‡Departamento de Estatística, Centro de Ciências Exatas e Tecnológicas, Universidade Federal de São Carlos, Caixa Postal 676, CEP: 13.565-905, São Carlos, São Paulo, Brasil, E-mail: *cirdemia.costa@gmail.com*

1 Introdução

Um dos objetivos nos estudos em análise de sobrevivência é conhecer o comportamento da função de sobrevivência, que é a probabilidade de um indivíduo (ou componente) sobreviver após um tempo pré-estabelecido e o comportamento da função de risco, que é a correspondente taxa de falha instantânea. A função de risco pode assumir diversas formas: constante, crescente, decrescente, unimodal ou em forma de banheira, porém, quando o comportamento da função de risco é não monótono, grande parte das distribuições usualmente conhecidas, tais como exponencial, Weibull entre outras, não são capazes de acomodar este tipo de taxa de falha, portanto a busca por modelos de probabilidade que acomodam essa variedade de formas de risco é a motivação de muitos trabalhos e deste também.

A distribuição exponencial é o modelo de probabibilidade mais popular para analisar dados de sobrevivência, mas tem a limitação de taxa de risco constante. Nas últimas décadas muitos autores têm proposto novas classes de distribuições, as quais são baseadas em modificações da distribuição Weibull para fornecer função de taxa de risco tendo forma de banheira. Entre essas, a distribuição Weibull exponenciada (Mudholkar et. al., 1995), a qual também exibe função de taxa de risco unimodal.

Em 1980, Ponnambalam Kumaraswamy propôs uma distribuição para aplicações em hidrologia, a distribuição Kumaraswamy (Kum). Tal distribuição é capaz de acomodar as taxas de falha monótonas e não monótonas. Está definida no intervalo (0,1), o que não é comum em estudos de análise de sobrevivência, pois, na maioria dos casos os tempos de falha assumem valores no conjunto dos reais positivos.

Se T é a variável aleatória contínua com distribuição Kumaraswamy, em que T pertence ao intervalo $(0,1)$, sua notação é $T \sim \text{Kum}(\varphi, \lambda)$. A sua função densidade de probabilidade (fdp) e função de distribuição acumulada (fda) são dadas, respectivamente por

$$f(t) = \varphi\lambda t^{\varphi-1}(1-t^{\varphi})^{\lambda-1}, \quad t \in (0,1) \tag{1}$$

e

$$F(t) = 1-(1-t^{\varphi})^{\lambda}, \tag{2}$$

sendo $\varphi > 0$ e $\lambda > 0$ os parâmetros de forma da distribuição.

A distribuição Kumaraswamy está relacionada à distribuição Beta da seguinte maneira: se $Y \sim \text{Beta}(1,\lambda)$, sua φ-ésima raiz ($T = \sqrt[\varphi]{Y}$) tem distribuição $T \sim \text{Kum}(\varphi,\lambda)$. A distribuição Kum é tão versátil quanto a distribuição Beta, porém mais simples especialmente em estudos de simulação, uma vez que ela assume uma forma fechada e simples para a fdp e para a fda.

Mais recentemente, baseada na distribuição Kumaraswamy, Cordeiro e de Castro (2009) propuseram uma nova família de distribuições, a Kumaraswamy-Generalizada (Kum-G). Ela é flexível e contém distribuições com funções de risco unimodal e em forma de banheira, como mostrado por Pascoal *et al*. (2011),

além de ter como casos especiais outras distribuições: a normal, Weibull, gama, Gumbel e Gaussiana inversa. O domínio da nova distribuição será o intervalo em que os casos particulares estão definidos, por exemplo, se o caso particular for a Weibull, a nova distribuição será Kum-Weibull e o seu domínio será os reais positivos, o qual é o domínio da distribuição Weibull.

Definição: Seja $G(t)$ a função de distribuição acumulada de uma variável aleatória T qualquer. Então, a função distribuição acumulada da distribuição Kum-G é estabelecida por

$$F(t) = 1 - \left[1 - G(t)^{\lambda}\right]^{\varphi}, \tag{3}$$

em que $\lambda > 0$ e $\varphi > 0$ são os dois parâmetros adicionais para a distribuição $F(t)$. Seja $g(t) = \dfrac{dG(t)}{dt}$, a função de densidade de probabilidade correspondente é

$$f(t) = \lambda\varphi g(t) G(t)^{\lambda-1} \left[1 - G(t)^{\lambda}\right]^{\varphi-1}. \tag{4}$$

Neste artigo combinou-se os trabalhos de Kumaraswamy (1980) e Cordeiro e de Castro (2011) para estudar as propriedades de um novo modelo e aplicar a dados de sobrevivência, o chamado modelo Kumaraswamy-Exponencial (Kum-Exp).

A organização deste trabalho está como segue. A Distribuição Kumaraswamy-Exponencial e suas propriedades são apresentadas na sessão 2. Na sessão 3, são apresentados o estimador de máxima verossimilhança dos parâmetros do modelo, estudo de simulação e aplicação a dados reais. A abordagem bayesiana é dada na sessão 4, bem como aplicação a dados reais. O trabalho é finalizado com as considerações finais do desenvolvimento deste estudo.

2 Distribuição Kumaraswamy-Exponencial

Para estabelecer a distribuição Kum-Exp como um caso especial da Kum-G, a função de distribuição acumulada da distribuição exponencial com parâmetro α é $G(t) = 1 - e^{-\alpha t}$ para $t > 0$. Correspondentemente a função densidade de probabilidade e a função distribuição acumulada da distribuição Kumaraswamy-Exponencial ($T \sim$ Kum-Exp$(\varphi, \lambda, \alpha)$) são

$$g(t) = \varphi\lambda\alpha \mathrm{e}^{-\alpha t}(1 - \mathrm{e}^{-\alpha t})^{\lambda-1}[1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi-1}. \tag{5}$$

e

$$F(t) = 1 - [1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}. \tag{6}$$

Os parâmetros $\varphi > 0$, $\lambda > 0$ e $\alpha > 0$, são parâmetros de forma. Quando os parâmetros assumem os valores ($\varphi = 1$,$\lambda = 1$), ($\alpha = 1$,$\lambda = 1$) e ($\lambda = 1$),

tem-se casos particulares da Kum-Exp, em que todos chegam na distribuição Exponencial, diferindo apenas nos parâmetros, sendo, respectivamente, $Exp(\alpha)$, $Exp(\varphi)$ e $Exp(\alpha\varphi)$.

A função de sobrevivência e a função de risco, respectivamente, são dadas por

$$S(t) = [1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi} \tag{7}$$

e

$$h(t) = \frac{\varphi\lambda\alpha\mathrm{e}^{-\alpha t}(1 - \mathrm{e}^{-\alpha t})^{\lambda-1}}{1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}}. \tag{8}$$

A Figura 1 mostra as representações gráficas da fdp, da função de sobrevivência e de risco para alguns valores dos parâmetros. Esses gráficos mostram a grande flexibilidade da distribuição Kum-Exp.

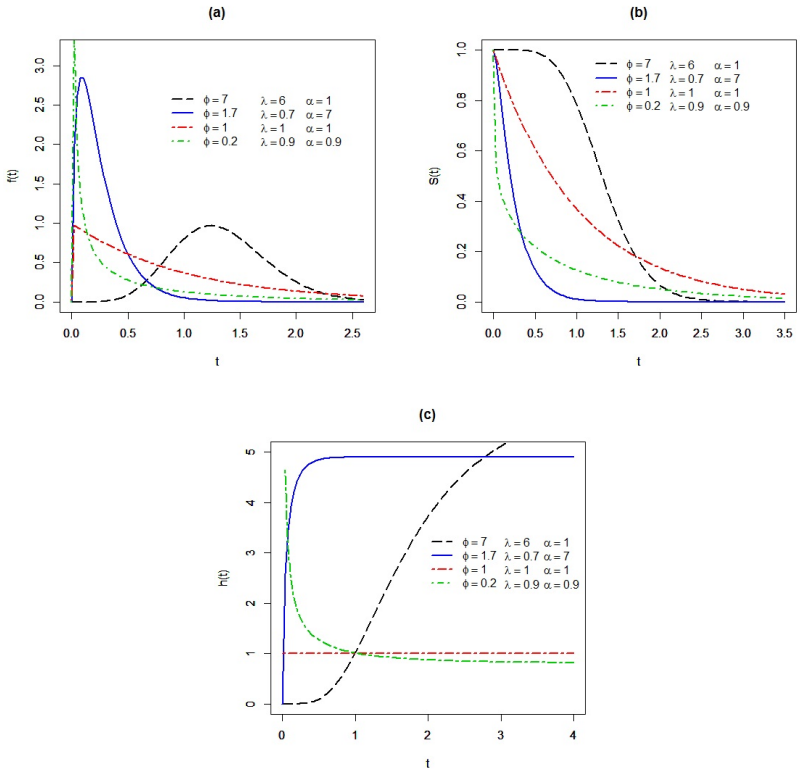

Figura 1: Gráficos das funções de densidade (a), sobrevivência (b) e risco (c) da distribuição Kum-Exp.

Interpretação física: seja um sistema formado com φ componentes independentes, cada um destes componentes é formado por λ sub-componentes independentes. O sistema falha se qualquer dos φ componentes falhar e um dos componentes falha se todos os λ sub-componentes falharem. Também, $T_{j1}, ..., T_{j\lambda}$ representam os tempos de sobrevivência dos sub-componentes do j-ésimo componente, onde $j = 1, ..., \varphi$, todos eles tendo a mesma fda $G(t)$, que é a fda da distribuição exponencial. Suponha T_k o tempo de sobrevivência do componente k, para $k = 1, ..., \varphi$, e T o tempo de sobrevivência de todo o sistema. Dessa forma, a distribuição Kum-Exp pode ser interpretada como a distribuição do tempo de falha do sistema inteiro.

A Esperança e a Variância são comumente utilizadas para expressar uma medida de tendência central e de variabilidade dos dados, respectivamente. O Método dos Momentos é uma das maneiras de calcular tais medidas, Cordeiro *et al*. (2009) propuseram o cálculo para os momentos da distribuição Kum-Exp dado pela seguinte fórmula:

$$E(T^n) = n\lambda^n \sum_{i,j=1}^{\infty} W_i \frac{-1^{n+j} \binom{\lambda(i+1)-1}{j}}{(j+1)^{n+1}}, \tag{9}$$

em que $T > 0$ é a variável aleatória; $W_i = (-1)^i \lambda\varphi \binom{\varphi-1}{i}$; $j = 1, \ldots, \varphi$; $i = 1, \ldots, \lambda$; sendo que φ é um valor positivo natural que indica o número de componentes independentes e λ é um valor positivo natural que indica o número de sub-componentes independentes.

A função quantílica, utilizada em simulações, é dada por

$$\begin{aligned} F^{-1}(t) &= G^{-1}\left\{\left[1-(1-t)^{1/\varphi}\right]^{1/\lambda}\right\} \\ &= -\frac{1}{\alpha}\log\left\{\left[1-(1-t)^{1/\varphi}\right]^{1/\lambda}\right\}, \end{aligned} \tag{10}$$

assim, basta simular valores de uma variável aleatória Uniforme(0,1), substituir em t na função quantílica e teremos valores simulados da distribuição Kum-Exp$(\varphi, \lambda, \alpha)$.

3 Inferência

Para modelar o tempo de vida de indivíduos é necessário estimar as funções de sobrevivência e de risco. Devido a propriedade de invariância do estimador de máxima verossimilhança, para estimar as funções (7) e (8) da distribuição Kum-exp, basta estimar o vetor de parâmetros $\boldsymbol{\theta} = (\varphi, \lambda, \alpha)$.

Considerando dados de sobrevivência, a partir dos tempos observados e do

vetor de parâmetros, a função de verossimilhança da Kum-Exp é dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}} \right]^{\delta_i} \left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi} \right] \tag{11}$$

na qual δ_i é a indicadora de censura (1 indica falha e 0 indica censura).

O logarítmo da função de verossimilhança é

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \delta_i \log \left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}} \right] \\ &\quad + \sum_{i=1}^{n} \varphi \log\left[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}\right]. \end{aligned} \tag{12}$$

Os estimadores de máxima verossimilhança para os três parâmetros são encontrados a partir da resolução da equação de verossimilhança

$$U(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0. \tag{13}$$

Resolvendo a equação de verossimilhança (13) para φ, o seu EMV é

$$\hat{\varphi} = -\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} \log[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}]}. \tag{14}$$

Embora tenha sido possível encontrar uma expressão para o parâmetro φ, ele depende dos demais parâmetros λ e α, e seus respectivos estimadores não puderam ser encontrados de forma analítica. Portanto, a estimativa por máxima verossimilhança para os três parâmetros φ, λ e α devem ser obtidas via métodos numéricos.

3.1 Estudo de Simulação

A fim de verificar as propriedades frequentistas, um estudo de simulação foi realizado supondo que o tempo de vida dos indivíduos em risco segue uma distribuição Kum-Exp. Para iniciar o processo de simulação, primeiramente os parâmetros foram fixados aleatoriamente da seguinte maneira: $\varphi = 10$, $\lambda = 3$ e $\alpha = 1,8$. Foram realizados dois estudos, um deles considerando a presença de censura e o outro não. Para o estudo com censuras, supôs-se que os tempos de censura seguem, também, a distribuição Kum-Exp, para os parâmetros fixados $\varphi = 14$, $\lambda = 1$ e $\alpha = 1,6$.

Utilizou-se três tamanhos amostrais: $n = 100$, $n = 200$ e $n = 500$. Foram geradas 1.000 réplicas, sendo que em cada etapa, foram obtidas as estimativas de máxima verossimilhança dos parâmetros, a partir da rotina "mle" do software livre "R". Foram atribuídos como valores iniciais para o processo de otimização os valores dos parâmetros pré-fixados para os tempos de falha.

Em cada simulação foi obtido o Erro Quadrático Médio (EQM) de cada estimativa e a amplitude do intervalo de confiança de 95%, baseado na teoria assintótica. O EQM foi calculado da seguinte maneira:

$$EQM = \sum_{i=1}^{n} \frac{(\theta_i - \widehat{\theta}_i)^2}{d},$$

em que θ_i é o valor fixado de cada parâmetro e $\widehat{\theta}_i$ é a média aritmética das d = 1000 réplicas para cada parâmetro estimado, isto é, para $\widehat{\varphi}$, $\widehat{\lambda}$ e $\widehat{\alpha}$.

O intervalo com 95% de confiança (IC(95)) foi calculado da seguinte forma:

$$IC(1-\gamma) = \widehat{\theta}_i \ \pm \ \sigma_i \ Z_{\gamma/2},$$

em que $Z_{\gamma/2}$ é o quantil $\gamma/2$ da distribuição Normal e σ_i é o desvio padrão de cada parâmetro. O valor de γ utilizado foi de 0,05.

O seguinte algorítmo foi utilizado:

1. Fixar os valores dos parâmetros;
2. Gerar $u_i \sim U(0,1)$;
3. Gerar y_i, tal que $y_i = -\dfrac{\log(1-(1-(1-u_i)^{(1/\lambda)})^{(1/\varphi)})}{\alpha}$, utilizando os parâmetros do tempo de falha;
4. Gerar $v_i \sim U(0,1)$;
5. Gerar x_i, tal que $x_i = -\dfrac{\log\left(1-(1-(1-v_i)^{(1/\lambda)})^{(1/\varphi)}\right)}{\alpha}$, utilizando os parâmetros do tempo de censura;
6. O tempo do i-ésimo indivíduo é $t_i = min(y_i, x_i)$;
7. Se $y_i \leq x_i$ faça $\delta_i = 1$, caso contrário, $\delta_i = 0$;
8. Repetir o processo do item 2 ao item 7 n vezes;
9. Repetir o item 2 a 8 até obter $'d'$ réplicas.

para $d = 1000$ e $n = 100, 200, 500$. Observação: para o conjunto de dados sem censura, pular os itens de 4 a 7 e considere $\delta_i = 1$.

As estimativas obtidas com dados censurados e completos estão dispostas na Tabela 1 e Tabela 2 respectivamente. Onde observamos que as médias das estimativas não foram muito afetadas pelo aumento no tamanho da amostra, mesmo com amostra pequena a média das estimativas ficaram próximas dos valores verdadeiros. O EQM e Ampliude do intevalo decresce à medida que o tamanho amostral cresce.

A Figura 2 mostra o comportamento do EQM e da amplitude do intervalo de confiança das simulações sem censura e a Figura 3 mostra o comportamento dos mesmos, porém com os dados completos.

Tabela 1: Estimativas da Simulação - Dados com Censura

n	Parâmetros	Estimativas	EQM	Amplitude do IC
100	φ	11,462	33,261	2,151
	λ	3,566	4,865	3,962
	α	1,887	0,311	2,073
200	φ	10,256	2,153	1,461
	λ	3,142	0,928	2,717
	α	1,811	0,028	1,435
500	φ	10,437	3,837	0,899
	λ	3,12	0,919	1,67
	α	1,84	0,089	0,883

Tabela 2: Estimativas da Simulação - Dados Completos

n	Parâmetros	Estimativas	EQM	Amplitude do IC
100	φ	11,541	28,945	2,003
	λ	3,358	3,008	3,430
	α	1,915	0,304	1,858
200	φ	10,171	2,337	1,352
	λ	3,179	1,151	2,402
	α	1,799	0,033	1,306
500	φ	10,335	3,255	0,835
	λ	3,102	0,649	1,471
	α	1,828	0,074	0,802

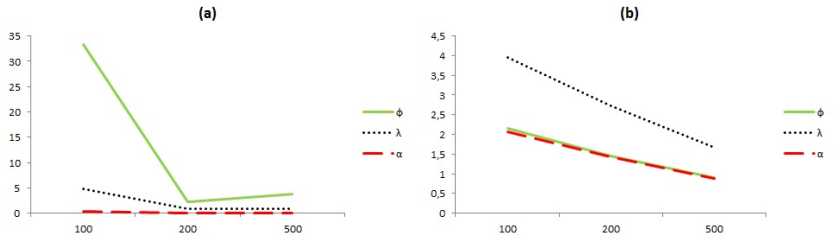

Figura 2: Comportamento do EQM (a) e da amplitude do intervalo de confiança (b) das simulações sem censura.

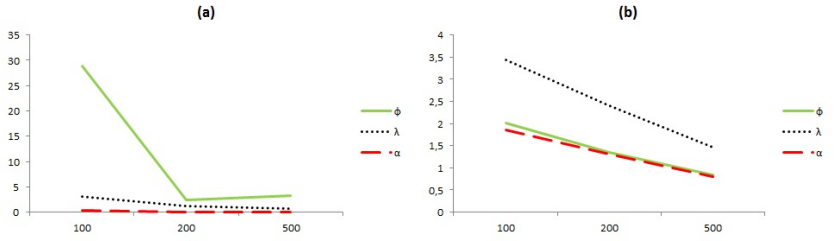

Figura 3: Comportamento do EQM (a) e da amplitude do intervalo de confiança (b) das simulações com dados completos.

3.2 Aplicação

Nesta seção será considerado dois conjuntos de dados reais para ilustrar a metodologia proposta.

3.2.1 Exemplo 1: Dados de células de redução de alumínio

Este conjunto de dados refere-se à tempos de falha de 20 células de redução de alumínio, em anos. Os dados podem ser encontrados em Whitmore (1983). Neste conjunto de dados existem 3 censuras. O interesse é modelar a função de sobrevivência destes tempos, utilizando o modelo Kum-Exp.

Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo e os intervalos de confiança de 95% são apresentados na Tabela 3.

Tabela 3: Estimativas dos parâmetros - Exemplo 1

Parâmetros	EMV	IC 95%
λ	$4,7283$	$(2,4576; 6,9991)$
φ	$2,1076$	$(-4,4945; 8,7098)$
α	$0,3582$	$(-3,7096; 4,4261)$

A Figura 4 apresenta os gráficos do ajuste da curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelo modelo. Pode-se observar que as curvas estão bem próximas, tendo um indicativo de que este modelo é adequado para os dados.

3.2.2 Exemplo 2: Dados de tempo até a morte de ratos por câncer vaginal

Este conjunto de dados refere-se à um grupo de ratos que foi exposto ao câncer, e foi registrado o tempo, em anos, até a morte por câncer vaginal. Há a presença

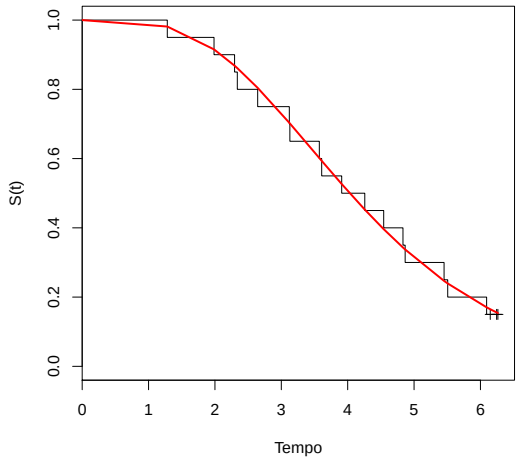

Figura 4: Curva de Kaplan-Meier juntamente com a função de sobrevivência estimada dos dados.

de 2 censuras. Em Pike (1966) encontram-se os dados de dois grupos de ratos, mas foi analisado apenas o primeiro grupo. Novamente, o interesse é modelar a função de sobrevivência destes tempos.

Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo e os desvios padrões são apresentados na Tabela 4.

Tabela 4: Estimativas dos parâmetros - Exemplo 2

Parâmetros	EMV	IC 95%
λ	$40,1070$	$(36,9590; 43,2551)$
φ	$4,6809$	$(0,9592; 8,4027)$
α	$5,1137$	$(3,5361; 6,6913)$

A Figura 5 apresenta os gráficos do ajuste da curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelo modelo. Observa-se que as curvas estão bem próximas, sendo um indicativo de que este modelo é adequado para os dados.

4 Uma Análise Bayesiana

Sob o enfoque bayesiano, pode-se expressar a incerteza a respeito do parâmetro antes de observar os dados, utilizando uma distribuição a priori para os parâme-

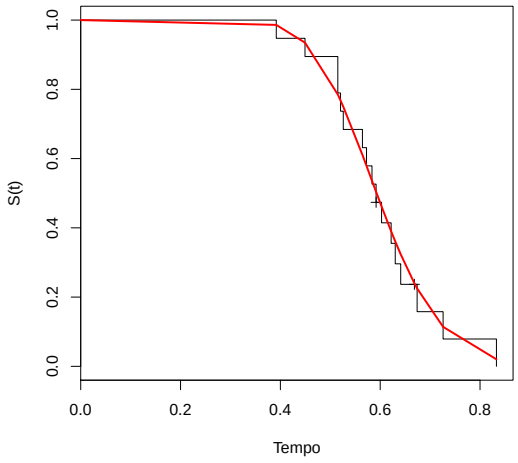

Figura 5: Curva de Kaplan-Meier juntamente com a função de sobrevivência estimada dos dados.

tros, enquanto que a distribuição a posteriori une a informação contida na verossimilhança com a distribuição a priori e, basicamente, as estimativas bayesianas são construídas a partir dessa distribuição a posteriori.

Para se estimar de forma bayesiana, deve-se construir a distribuição a posteriori que, pelo Teorema de Bayes, é dada por

$$\pi(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta})\pi(\boldsymbol{\theta}), \tag{15}$$

onde $\boldsymbol{\theta} = (\varphi, \lambda, \alpha)$ é o conjunto de parâmetros do modelo, $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ os tempos observados, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ a indicadora de censura e $L(\boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{\delta})$ é a função de verossimilhança do modelo.

Considerando que os parâmetros são independentes, então a priori de $\boldsymbol{\theta}$ é dada por

$$\pi(\boldsymbol{\theta}) = \pi(\lambda|\alpha_1, \beta_1)\pi(\varphi|\alpha_2, \beta_2)\pi(\alpha|\alpha_3, \beta_3), \tag{16}$$

onde α_1, β_1, α_2, β_2, α_3 e β_3 são os hiperparâmetros relacionados a $\boldsymbol{\theta}$.

Como o parâmetro λ assume valores positivos, admitiu-se que λ tem distribuição a priori Gama(α_1, β_1) e, da mesma forma, admitiu-se que φ tem distribuição a priori Gama(α_2, β_2) e que α tem distribuição a priori Gama(α_3, β_3).

Combinando a função de verossimilhança (11) e a densidade a priori de $\boldsymbol{\theta}$

(16), obtém-se a densidade a posteriori, dada por

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) &\propto L(\boldsymbol{\theta};\boldsymbol{t},\boldsymbol{\delta})\pi(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\times \pi(\lambda|\alpha_1,\beta_1)\pi(\varphi|\alpha_2,\beta_2)\pi(\alpha|\alpha_3,\beta_3). \qquad (17)\end{aligned}$$

Integrando 17 com relação a cada um dos parâmetros, são obtidas as densidades marginais a posteriori de cada um dos parâmetros, mas estas integrais não são analiticamente calculáveis. Uma alternativa é fazer o uso das densidades condicionais completas de todos os parâmetros, dadas por

$$\begin{aligned}\pi(\lambda|\varphi,\alpha,\boldsymbol{t},\boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\times \lambda^{\alpha_1-1}e^{-\beta_1\lambda}\end{aligned}$$

$$\begin{aligned}\pi(\varphi|\lambda,\alpha,\boldsymbol{t},\boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\times \varphi^{\alpha_2-1}e^{-\beta_2\varphi}\end{aligned}$$

$$\begin{aligned}\pi(\alpha|\lambda,\varphi,\boldsymbol{t},\boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\times \alpha^{\alpha_3-1}e^{-\beta_3\alpha}\end{aligned}$$

As densidades condicionais não apresentam nenhuma distribuição conhecida, então pode ser feito o uso do algoritmo de Metropolis-Hastings para gerar valores de λ, φ e α. Tal algoritmo permite simular amostras de distribuições conjuntas complexas, utilizando as distribuições condicionais completas dos parâmetros desconhecidos.

Para verificar a convergência do algoritmo de Metropolis-Hastings, pode-se fazer o uso de técnicas gráficas e o uso de algum método numérico. O utilizado foi de Gelman-Rubin (Gelman e Rubin, 1992), que está implementado no sistema R, juntamente com a análise gráfica.

4.1 Aplicação

4.1.1 Exemplo 1: Dados de células de redução de alumínio

A aplicação do conjunto de dados referente à tempos de falha de 20 células de redução de alumínio foi refeita utilizando a abordagem bayesiana. Para isso, considerou-se a distribuição Gama para as prioris com a média sendo o valor estimado dos parâmetros por máxima verossimilhança e variância igual a 10, da

seguinte forma: $\pi(\lambda) \sim \text{Gama}(2,2373; 0,473)$, $\pi(\varphi) \sim \text{Gama}(0,4452; 0,211)$ e $\pi(\alpha) \sim \text{Gama}(0,013; 0,036)$.

Para o uso do algoritmo Metropolis-Hastings, como λ, φ e α são positivos, foi gerado cada um deles através da distribuição exponencial, com a média sendo o valor simulado anteriormente, ou seja, geramos $\lambda_n \sim \text{Exponencial}(1/\lambda_a)$, $\varphi_n \sim \text{Exponencial}(1/\varphi_a)$ e $\alpha_n \sim \text{Exponencial}(1/\alpha_a)$, onde λ_n, φ_n e α_n são os novos valores gerados de cada um dos parâmetros e λ_a, φ_a e α_a são os valores gerados anteriormente de cada um dos parâmetros.

A partir disso, foram simulados 99.000 valores de cada um dos parâmetros, sendo que os 1.000 primeiros foram retirados. A fim de se ter dados não correlacionados, foram selecionados valores fazendo saltos de 20 em 20 valores, ficando finalmente com 4.900 valores simulados de cada um dos parâmetros.

A Tabela 5 apresenta, para cada parâmetro, a média a posteriori, mediana, desvio padrão, o intervalo de credibilidade 95% e o método numérico de Gelman-Rubim (R) que, sob convergência, o valor da estatística tende à 1. Considera-se convergência para valores de R menores ou iguais à $1,01$. Dessa forma, tem-se um indicativo de convergência.

Tabela 5: Resumos das distribuições a posteriori dos parâmetros do modelo

Parâmetros	Média	Mediana	DP	IC	R
λ	$5,5350$	$5,4880$	$0,7046$	$(4,2775; 7,0260)$	1
φ	$1,5250$	$1,4480$	$0,5902$	$(0,5831; 2,8714)$	1
α	$0,0921$	$0,0899$	$0,0255$	$(0,0488; 0,1495)$	$0,999$

A fim de verificar a convergência do algoritmo também foram utilizados métodos gráficos.

O primeiro método gráfico é através de histogramas. Basta dividir o conjunto de dados em três partes iguais e plotar a primeira e a última parte. Na Figura 6 os histogramas dos três parâmetros são apresentados, em que a cor azul é da primeira parte dos valores simulados, rosa é da segunda parte e lilás é quando os dois se encontram. Como nos três casos os histogramas ficaram muito próximos, este é um indicativo de que o algoritmo convergiu.

Da mesma forma, pode-se verificar a convergência através do gráfico das densidades da primeira e última parte dos valores. A Figura 7 mostra os gráficos das densidades dos três parâmetros. Novamente há um indicativo de convergência já que os gráficos estão muito próximos nos três casos.

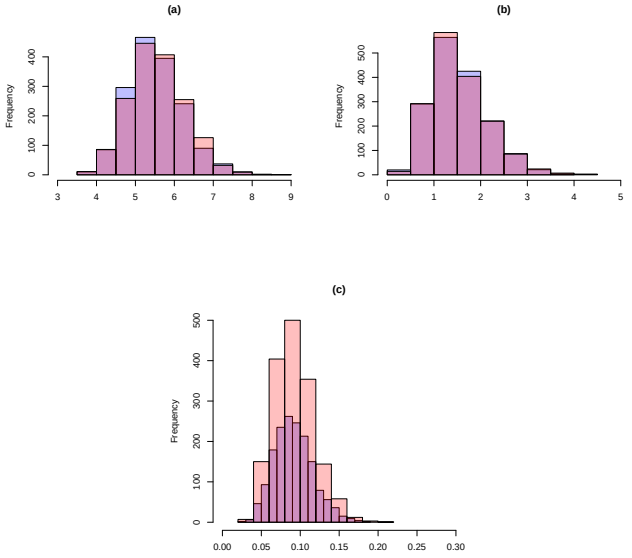

Figura 6: Gráficos dos histogramas do parâmetro λ (a), φ (b) e α (c).

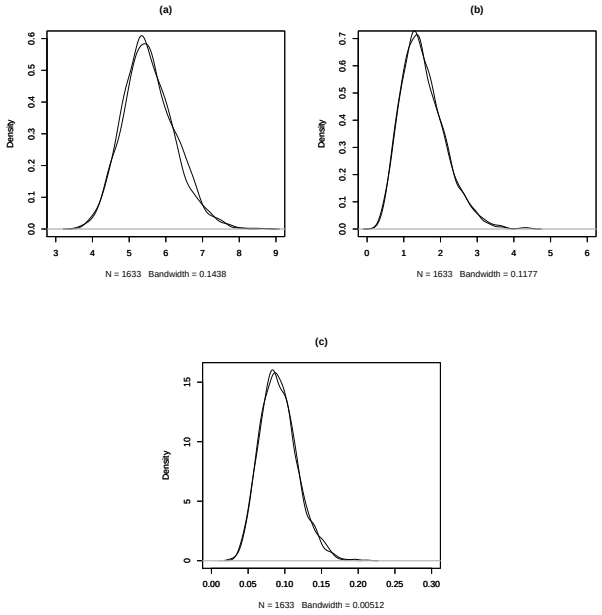

Figura 7: Gráficos de densidade do parâmetro λ (a), φ (b) e α (c).

Por fim, uma forma outra forma de verificar a convergência é através da sequencia dos valores simulados da primeira e última parte dos valores, uma das partes está em vermelho e a outra, em preto. A Figura 8 mostra que em todos os casos os gráficos estão muito próximos e sem nenhuma tendência, indicando novamente convergência do algoritmo.

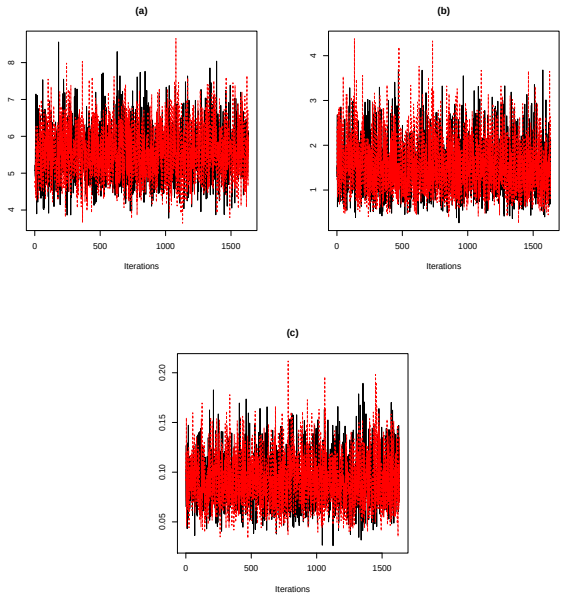

Figura 8: Gráficos da sequencia de valores do parâmetro λ (a), φ (b) e α (c).

4.1.2 Exemplo 2: Dados de tempo até a morte de ratos por câncer vaginal

Também foi refeita a aplicação do conjunto de dados referente ao tempo até a morte de ratos por câncer vaginal, utilizando a abordagem bayesiana. Novamente, considerou-se a distribuição Gama para as prioris com a média sendo o valor estimado dos parâmetros por máxima verossimilhança, da seguinte forma: $\pi(\lambda) \sim \text{Gama}(4; 0,1)$, $\pi(\varphi) \sim \text{Gama}(2.1911; 0,4681)$ e $\pi(\alpha) \sim \text{Gama}(2.6150; 0,5114)$.

Também foi gerado cada um dos parâmetros λ, φ e α através da distribuição exponencial, com a média sendo o valor simulado anteriormente, ou seja, foram gerados $\lambda_n \sim \text{Exponencial}(1/\lambda_a)$, $\varphi_n \sim \text{Exponencial}(1/\varphi_a)$ e $\alpha_n \sim \text{Exponencial}(1/\alpha_a)$.

A mesma quantia de valores dos parâmetros foi simulada, ficando novamente com 4.900 valores simulados de cada um dos parâmetros.

A Tabela 6 apresenta a média a posteriori, mediana, desvio padrão, o intervalo de credibilidade 95% e o método de Gelman-Rubin de para cada parâmetro.

Tabela 6: Resumos das distribuições a posteriori dos parâmetros do modelo

Parâmetros	Média	Mediana	DP	IC	R
λ	$39,5700$	$39,4100$	$4,4944$	$(31,2826; 48,9574)$	1
φ	$4,5380$	$4,5170$	$0,7084$	$(3,2512; 5.9830)$	1
α	$2,4580$	$2,4460$	$0,3401$	$(1,8409; 3,1503)$	1.01

A Figura 9 apresenta os histogramas dos três parâmetros, da mesma forma como feito no exemplo anterior. Como nos três casos os histogramas ficaram muito próximos, este é um indicativo de que o algoritmo convergiu.

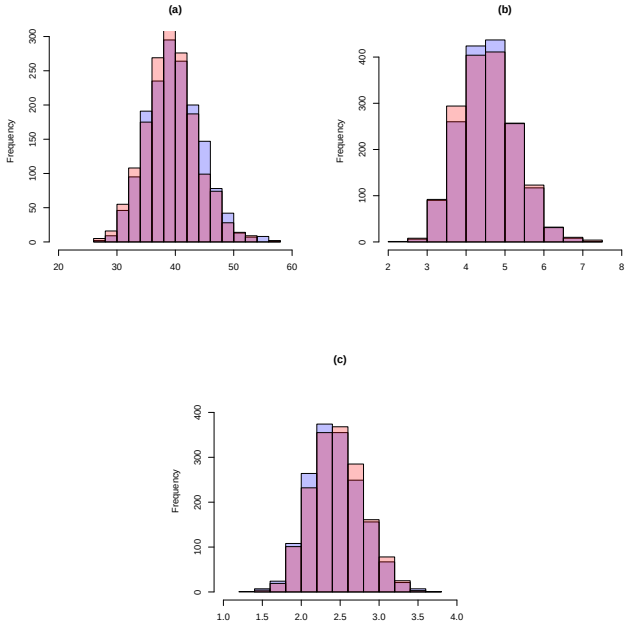

Figura 9: Gráficos dos histogramas do parâmetro λ (a), φ (b) e α (c).

A figura 10 exibe os gráficos das densidades dos três parâmetros. Tem-se novamente um indicativo de convergência já que os gráficos estão muito próximos nos três casos.

A Figura 11 mostra que em todos os casos os gráficos estão muito próximos e sem nenhuma tendência, indicando novamente convergência do algoritmo.

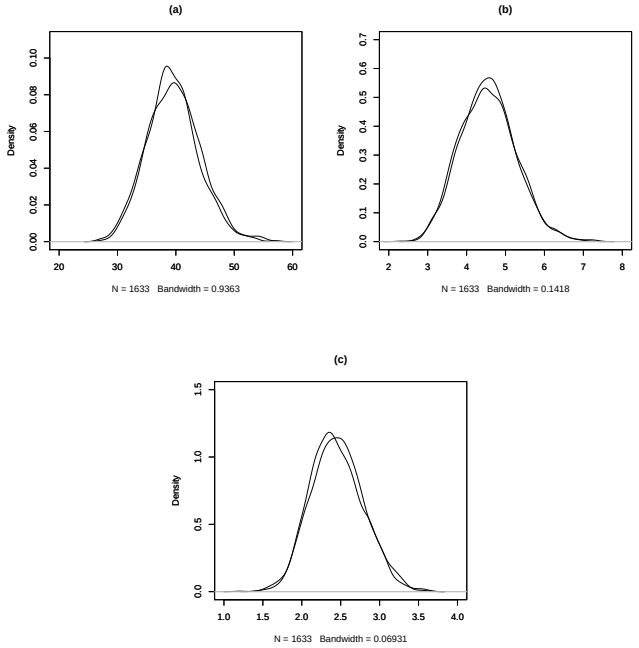

Figura 10: Gráficos de densidade do parâmetro λ (a), φ (b) e α (c).

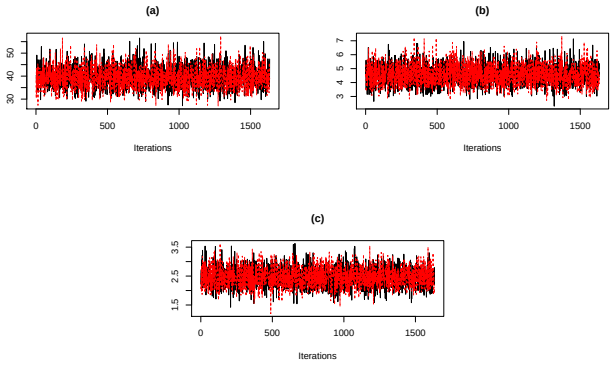

Figura 11: Gráfico da sequencia de valores dos parâmetro λ (a), φ (b) e α (c).

Considerações Finais

A distribuição de tempo de vida Kumaraswamy-Exponencial é proposta como uma simples extensão da distribuição exponencial, para a qual foram apresentadas as principais funções para análise de dados de sobrevivência. A estimação dos parâmetros da Kum-Exp foi feita pelo método de Máxima Verossimilhança e pela abordagem bayesiana. Oferecemos um tratamento matemático desta distribuição, incluindo as expressões explícitas para os momentos. A relevância e aplicabilidade prática do novo modelo são demonstradas nos conjuntos de dados reais.

No estudo de simulação, como esperado, viu-se que conforme o tamanho amostral aumenta, o valor das estimativas ficou próximo do valor real e o EQM e a amplitude do intervalo de confiança ficam cada vez menores, tanto para dados censurados, como para dados completos.

Referências

[1] Cordeiro, G. M. and de Castro, M., *A new family of generalized distributions*, Journal of Statistical Computation and Simulation, doi: 10.1080/0094965YY, 2009.

[2] Gelman, A. and Rubin, D., *Inference from iterative simulation using multiple sequences.*, Statistical Science, 7(4):457–472, 1992.

[3] Kumaraswamy, P., *A generalized probability density function for doublebounded random processes*, Journal of Hydrology, 46, 79-88, 1980.

[4] Mudholkar, G.S.; Srivastava, D.K.; Freimer, M., *The exponentiated Weibull family; a reanalysis of the bus motor failure data*, Technometrics 37 (4): 436–445, 1995.

[5] Pascoal, M. A. R., Ortega, E. M. M. and Cordeiro, G. M., *The Kumaraswamy generalized gamma distribution with application in survival analysis*, Statistical Methodology, 8, 411-433, 2011.

[6] Pike, M. C., *A method of analysis of a certain class of experiments in carcinogenesis*, Biometrics 22: 142–161, 1966.

[7] Whitmore, G. A., *A Regression Method for Censored Inverse-Gaussian Data*, The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Vol. 11, No. 4, pp. 305-315, Dec. 1983.